23

Role of Meta-analysis in Interpreting the Scientific Literature

Michael D. Jennions, Christopher J. Lortie, and Julia Koricheva

THE PROBLEM OF INFORMAL "EXPERT" ASSESSMENT OF RESEARCH FINDINGS

Scientists often deal with vast amounts of data, and the ability to summarize this information effectively is a major asset. Therefore, researchers make a considerable effort to acquire the necessary statistical skills to rigorously analyze each empirical data set that they collect. The same thoroughness should occur when writing up work for publication, which ideally requires synthesis of the scientific literature for each question that is answered (i.e. statistical test conducted) to place the results in context. This synthesis is a real challenge. One thing that makes it challenging for ecologists and evolutionary biologists to stay up to date with research findings is that, in so doing, they usually try to place their own results in a much broader context. This means that they often do not confine their frame of reference to studies of the same species, taxon, or ecosystem. There is, not unexpectedly, evidence that those working on a taxon studied by relatively few researchers are more likely to cite studies of other more popular study systems. For example, herpetologists more often cite studies of other classes of vertebrate than those studying mammals or birds cite reptilian studies (Bonnet et al. 2002; see also Taborsky 2009). In ecology and evolutionary biology, one reason for the inclusion of citations following specific research results is to illustrate the extent to which the author's results agree or disagree with the findings of others. Arguably, the most common way researchers assess the "level of agreement" is to consult reviews, including meta-analyses, to reach a general conclusion (e.g., most studies report a positive finding), which they can then claim their own study supports or contradicts.

The validity of any assessment of general findings depends on the rigor of the review used to inform this judgment. If the review is a meta-analysis, one can be fairly certain that the confidence intervals for the mean trend provide a reasonable quantitative summary of the available studies. If it is a narrative review, then the potential for a subjective bias on the part of the reviewer, including reliance on "expert opinion," which is surprisingly often flawed or erroneous (Surowiecki 2004), is more worrisome. The general conclusions drawn have a strong bearing on how research findings are presented (i.e., whether we describe results as refuting or supporting a general trend), and therefore have a major impact on the future direction of a researcher's own work and that of their colleagues (e.g., few researchers will ask a question if publications repeatedly state that we already know the answer).

In practice, reviews usually cover broad rather than specific questions. For example, it is easy to locate a review of the evolution of female mate choice, but finding one on the exact relationship between male advertisement call pitch and female mating preferences in frogs is more difficult. Consequently, attempts to summarize the findings of previous studies that tackle a specific question are rarely based on consultation of a quantitative, or even narrative, review. This can result in a highly idiosyncratic data synthesis process because each author must conduct their own separate review for every finding they wish to comment upon. This can lead to problems. Extrapolating from our own behavior and that of colleagues, many researchers tend to compile lists of papers that tackle a specific question. They then categorize these as reporting a significant positive or negative relationship, or failing to do either. Some researchers are disciplined and maintain spreadsheets of categorized publications, others are content with simply relying on their memory. If these lists were simply drawn upon to cite studies that have looked at the same question there would be no real concern. The problem, however, is that there is a temptation to tally up studies in an informal "vote count" (Chapter 1) and draw conclusions about general patterns. Aside from the problem that vote counting is a poor method to calculate trends, there is the underlying concern that the studies being tallied are a biased sample of those that have been conducted.

For many research questions (especially those subsidiary to the main focus of a study), it is probably fair to state that a good number of researchers simply consult a few recent papers and make a judgment as to the average outcome based on how many studies are cited as supporting or refuting a hypothesis. This raises a question: Are the cited studies a random sample of all those conducted? The answer in almost all cases is "no." Another commonly used shortcut to identify general trends is to simply accept at face value a published statement by a colleague that empirical support for a predicted outcome is rare or common. This is based on the assumption that he or she is an expert who has been more systematic in their review of the literature. This kind of copying can readily lead to a positive feedback loop (because authors rely on citations in the published literature that they themselves contribute to), and the emergence of fads and fashions with no link between the popular consensus and reality if the experts that initiate the feedback are wrong (Bikhchandani et al. 1992). Expert opinion is notoriously unreliable because many experts simply rely on their own biased and qualitative assessment of the literature (e.g., Antman et al. 1992).

A pragmatic approach to synthesizing the literature based on brief consultation of a subset of the published studies is understandable given time constraints. Even so, the potential for this to lead to misinformation should be readily apparent if citation and publication practices (which also inform "expert opinion") are associated with research findings (Chapter 14). First, judgment calls as to where the weight of evidence lies are often based on such factors as the relative ease with which one can recall studies that report positive and negative results. Unfortunately, humans have a propensity to recall certain events more readily than others; they are also prone to a range of other cognitive biases (Piatelli-Palmarini 1994). This can generate substantial memory bias as to the rate of occurrence of different types of events, such as pleasurable and painful experiences (Gilbert 2006). Although we are unaware of any formal investigation, it is worthwhile considering whether there is a greater likelihood of recalling a study that reports a highly significant relationship than one that fails to do so. Second, the findings of a study appear to influence the ease with which it is located in the literature, so careful attention must be paid to the potential for sampling bias. For example, the papers that appear to be more readily remembered and cited are those published in English in high-profile journals, and written by influential researchers, large research teams, or those working in the same country (Leimu and Koricheva 2005b, Wong and Kokko 2005). This is a potential problem because, due to publication bias,

some of these factors are associated with effect sizes and/or their variances (i.e., our statistical confidence in the estimate; Chapter 14). This could create a large discrepancy between the "conventional wisdom" of what has been shown by previous studies based on an informal summary of findings from more readily located or remembered work, and the outcome of a quantitative meta-analysis based on a well-defined sampling protocol (Chapters 3 to 5).

Some might argue that errors made during an informal assessment of a field are only a short-term problem because the truth will eventually prevail when a formal quantitative analysis is conducted. We think this attitude is counterproductive. In ecology and evolutionary biology, a decision on whether or not to test a hypothesis is largely dictated by the decisions of individual researchers, rather than panels or committees that determine policy and direct research. Unlike the sponsorship of research in some areas in the health sciences, natural science funding bodies do not require a formal meta-analysis as part of a grant application. They take it on good faith if the applicants state that they will work on a poorly studied or unresolved issue. At best, they seek confirmation from peer review and rely on "expert opinion" that, as already noted, is often flawed (Antman et al. 1992). This can lead to an enormous waste of resources if studies are designed to ask questions that have already been satisfactorily answered (although perhaps not in the same study system, but in a more general sense). In the medical sciences, for example, the use of cumulative meta-analysis has shown that costly large-scale trials have sometimes been conducted long after the efficacy of a treatment could be established through meta-analysis (Chapter 15). Conversely, systematic reviews that identify gaps in usable data (e.g., Stewart, Pullin, and Tyler 2007) or the potential for strong publication bias (Palmer 2000), reveal cases where phenomena that are widely described as well-established turn out to be unproven or disputable and therefore worthy of closer study. Anyone who has conducted a meta-analysis is familiar with the regularity with which well-known papers purporting to demonstrate a phenomenon either fail to do so, or do so without providing reliable information about the biologically relevant magnitude of the effect.

In the health sciences there are obvious ethical concerns about delays in the implementation of effective treatments and the associated problem of unnecessary research and badly presented research. For example, it was shown that studies continued to be conducted and patients assigned to control groups long after the efficacy of a treatment was demonstrated statistically (e.g., Fergusson et al. 2005 found that 52 more studies than necessary were conducted in one area of medicine; Chapter 15). This concern led the editors of the prestigious medical journal *The Lancet* to change the submission requirements for studies describing the results of new clinical trials. The authors are now asked to include a *"clear summary of previous research findings, and to explain how their trial's findings affect this summary." The journal also asks for the following: "The relation between existing and new evidence should be illustrated by direct reference to an existing systematic review and meta-analysis. When a systematic review or meta-analysis does not exist, authors are encouraged to do their own. If this is not possible, authors should describe in a structured way the qualitative association between their research and previous findings"* (Young and Horton 2005, 107). Although it seems unlikely to happen soon, a similar policy in ecological and evolutionary journals would be welcome.

Ideally, continually updated meta-analyses would be available for every research question posed for which sufficient information exists. Clearly this is not going to happen, but are there practical measures that can be implemented now? To start, we need to acknowledge that our potentially subjective and informal approach to synthesizing past work has created a scientific culture with an undue emphasis on *P*-values, simply because they allow studies to be designated as valuable, or ignored as inconclusive. An unfortunate byproduct of this practice is an unwarranted reliance on the information that can be extracted from an individual study,

especially one that reports a highly significant result. Specifically, we tend to treat studies with low *P*-values as being correct (or at least irrefutable), and discard studies with less clear-cut or nonsignificant findings when considering past work. This is a poor practice because a high percentage of positive (i.e., $P < 0.05$) results might actually be false; for a fascinating review, see Ioannidis 2005c, and see Ioannidis 2008 for a review of the related issue of inflated estimates. Ioannidis (2005c) derives simple equations to predict the poststudy probability that a statistically significant result is true (i.e., that the actual effect differs from the null value). A key equation is that a positive result is more likely true than false when $(1-\beta)R > \alpha$. Here $1-\beta$ = power (i.e., 1 − type II error rate), R = ratio of true effects to no effects in the field of study, and α is the type I error rate (usually 0.05). Consequently, the proportion of false positive results is higher when sample sizes are small and true effects are weak (as these decrease statistical power). Ioannidis (2005c) similarly presents equations to show how both a publication bias toward positive results (e.g., due to multiple testing and selective reporting) and increased numbers of researchers testing the same question further increase the proportion of false positive results. These insights should be of particular concern in ecology and evolutionary biology because in these fields (1) studies often have low sample sizes (Jennions and Møller 2003); (2) true effect sizes are often small (Møller and Jennions 2002); (3) numerous relationships are usually tested because studies are often exploratory so *R* can be small (e.g., whether extinction risk is related to body size, population size, sexual dimorphism, diet, clutch size, and so on); (4) statistical approaches are not formalized even when data sets have identical structures, which encourages "statistical fishing"; (5) several outcome variables are usually examined (e.g., effects of elevated CO_2 on growth of different plant parts, osmoregulation, or rates of photosynthesis) all of which can, post hoc, be described as important; and (6) in some "hot topic areas," such as climate change, numerous research teams are each testing the same hypotheses (e.g.. whether the onset of breeding moved forward in species X in the last 40 years).

What can be done to help researchers better summarize developments in ecology and evolution? Readers of Chapter 1 will hopefully agree that a modern meta-analysis offers a research summary that is superior to vote counting based on critical *P*-values or to (worse still) reliance on statements about the strength of relationships taken from narrative reviews. Of course, when a published meta-analysis is unavailable a researcher must rely on their own synthesis of the field. Papers report *P*-values rather than effect sizes, so the path of least resistance is to filter primary studies through the sieve of threshold *P*-values. The solution is two-pronged, and will ensure a more mature approach to the assessment of statistics and create conditions that should make quantitative reviews easier to conduct. First, ecological and evolutionary journals must encourage the reporting of a wider spectrum of outcomes (e.g., 95% confidence intervals for effect sizes, investigation of sources of heterogeneity among studies), rather than *P*-values (if the 95% CI does not overlap the null value, the reader immediately knows that $P < 0.05$ anyway). Second, researchers must learn to evaluate studies in terms of this wider spectrum of outcomes, rather than *P*-values. Figure 23.1 illustrates the interacting impact of a range of measures on a meta-analysis, including not only the effect estimate, but also sample sizes and the corresponding confidence intervals.

The approach we take in the remainder of this chapter is motivated by our own real world experiences as ecologists and evolutionary biologists who have conducted meta-analyses. Collating data for a meta-analysis invariably leads to a shift from a worldview where the focus is on seeking the truth based on single "perfect" studies ("textbook examples," Chapter 1) to one that regards the literature as a population of studies, each with one or more effect sizes. These effect sizes are estimates of the "true" effect size so that their pooled magnitude, variance, and heterogeneity become the real focus of attention. We therefore begin with a brief review

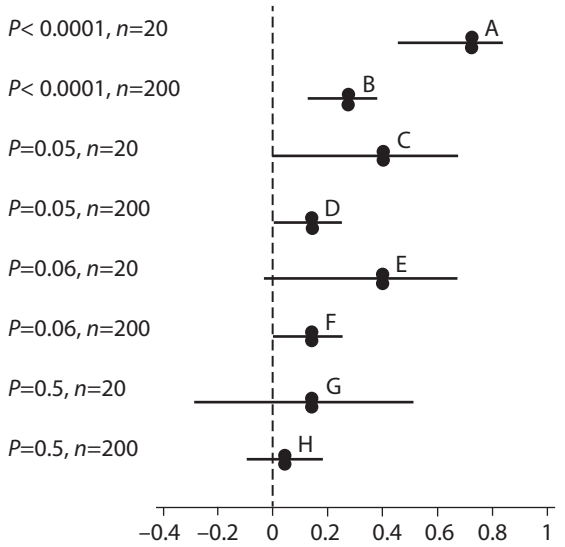

Figure 23.1. The relationship between *P*-value, sample size, estimated mean effect size (in this case *r*, Pearson's correlation coefficient), and confidence intervals showing the effect of a change in sample size with the same level of significance (modified from Nakagawa and Cuthill 2007).

of why effect sizes and their variances (usually expressed as confidence intervals) are more informative than *P*-values. We then discuss how meta-analysis promotes "effective thinking" (Nakagawa and Cuthill 2007) that can change approaches to several commonplace problems. Specifically, we address the issues of (1) exemplar studies versus average trends, (2) resolving "conflict" between specific studies, (3) presenting results, (4) deciding on the level at which to replicate studies, (5) understanding the constraints imposed by low statistical power, and (6) asking broad-scale questions that cannot be resolved in a single study.

In this chapter, we focus on estimating effect sizes as a key outcome of meta-analysis, but acknowledge that other outcomes might be of more interest in other situations. These could include, for example, comparisons between effect sizes, hypothesis testing, evaluation of moderators of effect sizes, and identifying other sources of heterogeneity between studies. However, we would argue that the points we make in the context of estimating effect sizes apply more generally to this wider meta-analysis spectrum. We should note that, for brevity, we often do not distinguish between parameter estimation and hypothesis testing. The standard null hypothesis in most areas of ecology and evolutionary biology is that a measured parameter (the "effect size" in a meta-analysis) has a mean value of zero or a nonzero theoretically predicted value (e.g., for allometric scaling, see Chapter 24). In some cases, however, parameters are estimated without being used to test a formal hypothesis (e.g., the annual rate of decline in coral cover).

EFFECT SIZES VERSUS *P*-VALUES

Ecologists and evolutionary biologists have continually been encouraged to switch from a frequentist approach of null hypothesis significance testing (i.e., whether *P*-values cross a threshold value like 0.05) toward other statistical approaches in order to summarize their research findings (e.g., Fernandez-Duque 1997, Johnson 1999, Stoehr 1999, Jennions and Møller 2003, Nakagawa 2004). These calls have largely gone unheeded. For example, adoption of

Bayesian approaches has been erratic and has only occurred in some subdisciplines (Garamszegi et al. 2009). This is probably because established biologists lack the time to master unfamiliar statistical theory, and user-friendly software is limited (although more recent introductory textbooks for ecologists might facilitate a shift; e.g., McCarthy 2007). In contrast, the suggestion that ecologists and evolutionary biologists summarize data by presenting effect sizes and their confidence intervals is a relatively undemanding request. Suitable software exists and effect sizes are readily conceptualized as "sample size corrected" versions of familiar test statistics such as F or t. In other words, these test statistics are formulated by combining effect and sample sizes (Rosenthal 1994; Chapters 6 and 7).

Why then are data in almost every ecological and evolutionary paper still summarized using P-values? Researchers are perhaps unaware of the benefits that reporting effect sizes offer. This is understandable when one considers that the benefits often accrue to the scientific community (e.g., the ability to conduct a meta-analysis) rather than to the individual author, who might even pay a cost. For example, if $P < 0.01$ and the average referee currently takes this as a sign of a "clean result," why reduce the chance of publication by reminding referees that the estimated effect size is small or the confidence interval wide? More practically, researchers often do not know how to calculate effect sizes and their variances. This might be easy for many standard test statistics, but it is trickier for others (Chapters 6 and 7; Nakagawa and Cuthill 2007). Reviewers are not known for their leniency toward those who say: "I did the right thing and used approach X, except when it was really tricky and would have taken me ages to work out how to do it." Editors would, however, be doing the fields of ecology and evolutionary biology a service if they at least required authors to provide effect sizes for a prescribed set of simple statistical tests such as unpaired t-tests or F-values from one-way ANOVAs. Of course, the growing use of meta-analysis (Chapter 1) might, by itself, stimulate a change in how statistical tests are reported. Experience suggests, however, that reporting effect sizes in primary empirical studies will not become widespread until journals make it a prerequisite. Otherwise, it merely adds another chore to the publication process with no obvious reward to the researcher.

So why report effect sizes? Presenting effect sizes and their confidence intervals allows for better interpretation of data than examination of P-values. The standard null hypothesis is that no relationships between variables or differences among groups exist. P-values simply indicate the likelihood that a relationship or difference as or more extreme than that observed will occur if the null hypothesis is true. In short, a P-value only tells us whether the 95% confidence interval for an effect size includes the null expectation. This creates a dichotomy that can be handy, but can mislead the unwary reader because it ignores both the width and central location of the confidence interval. These two extra pieces of information can lead to a radical reassessment of a result initially interpreted solely on the basis of P-values. For example, many researchers might agree with the statement that a relationship in their field of study is important if they are only told that $P < 0.01$. They would, however, probably revise their opinion if informed that the 95% CI is $r = 0.03$ to 0.16 (e.g., when $n \approx 1000$) because the fairly narrow confidence interval means that the true correlation is probably weak. Similarly, if $P < 0.01$, even if the estimated effect is large, say $r = 0.49$, researchers would probably not describe the relationship as strong when told that the 95% CI was $r = 0.12$ to 0.74 (e.g., when $n \approx 25$) because the estimate is so imprecise. Finally, even when a result is nonsignificant, any biological interpretation should be influenced by the width of the confidence interval. If large (e.g., $r = -0.30$ to 0.45), we recognize that the data are inconclusive; if small we can conclude that an effect is either weak or absent. In sum, P-values are only biologically meaningful when sample sizes are taken into account. (Even then, information on the direction of the relationship is essential. Though a seemingly obvious point, this is information that often goes unreported when presenting

nonsignificant results (e.g., Cassey et al. 2004)). Graphical presentation of effect sizes is an efficient way to reveal the limits of using *P*-values (see Fig. 23.1 taken from Nakagawa and Cuthill 2007).

Are we exaggerating the problem? To drive home the danger of focusing on *P*-values it is worth answering a simple question: How often have you responded to a colleague's query about your latest study by saying something like, "great news, we got a significant result," without any mention of sample size and thus the effect size or its confidence interval? Most of us, if honest, can only reply, "I do that all the time." This illustrates the point that, in practice, we fail to distinguish between studies A to D, and often lump together studies E to G in Figure 23.1 when we present our research findings to others. The take-home message gleaned from casual conversations about other people's studies (whether it was a positive or negative result) is very similar to the information we retain when we finish reading a paper and rely on *P*-values to summarize what was reported. For those interested in an extended but readable account of the benefit of using effect sizes rather than *P*-values for ecologists we recommend Nakagawa and Cuthill (2007).

WHAT IS SO GREAT ABOUT YOUR STUDY?

It is useful to think of the practice of science as involving two competing tasks. First, to replicate precisely any study that produces an exciting result to validate the finding. Second, to make broad generalizations that can predict or account for events across a wider range of circumstances (Chalmers 1999). In some respects, these conflicting demands mirror the extent to which different researchers emphasize individual *P*-values or the distribution of effect sizes. This is because study replication is often associated with confirming that a prior study with a positive result was valid (for a more detailed discussion of what constitutes successful replication, see Kelly 2006), while generalization is usually associated with estimating the mean and variance (and sometimes sources of heterogeneity) in effect sizes across a range of studies (i.e., meta-analysis). The tension between these demands can be acute for ecologists and evolutionary biologists because (1) biological systems that are nominally the same actually vary spatially and temporally, so it is unclear what constitutes satisfactory replication of a study; and (2) the variety of available study systems is immense so that controversies arise as to the appropriate level at which to seek generalities. For example, Palmer (2000) has described all studies that are replicated using different species or systems as "quasi replication." Some workers go so far as to argue that ecology is "a highly idiographic science best served by amassing a catalogue of case studies" (Simberloff 2006b, 921), while others prefer to seek generalities. An example of the debate this engenders is seen in the contrasting viewpoint of Gurevitch (2006) and Simberloff (2006a, 2006b) on how best to study the interactions between native and invasive species in testing for the existence of "invasional meltdowns."

Meta-analysis is a quantitative framework to answer questions about the mean strength and sources of variation in relationships across studies. However, because it has been underutilized by ecologists and evolutionary biologists, greater emphasis has been placed on the value of individual studies. Unfortunately, and perhaps as a result, a bizarre practice has crept into many areas of ecology and evolutionary biology; this is to identify a key study (often based on little more than a small *P*-value and/or a large response) and then extrapolate from its findings to a wider set of circumstances. In its crudest form this results in the deification of classic studies in textbooks, which are treated as exemplars of a wider reality ("textbook examples," Chapter 1). This approach might be defensible in disciplines where the phenomenon under study is relatively invariant, but it is almost certainly inappropriate in ecology and evolutionary biology

given the obvious biological differences among populations (with even more variation among species or higher taxa), the amount of background "environmental noise" in most studies, and large measurement error in many subdisciplines (e.g., evolutionary fitness is notoriously difficult to measure). The weakness of this approach to the literature is driven home if one calculates the probability that a single study will report the true mean effect size, even when we are only interested in knowing this for a single population under identical conditions. If one thinks in terms of a population of effect sizes measured with error, it is clearly improbable that this single study—especially given a publication bias favoring stronger findings—will report the true mean effect size. (For example, assuming a normal distribution, for any given study there is a 32% chance that its effect size will be more than one standard deviation from the true mean, and a 5% chance it is more than 1.96 standard deviation from the true mean.)

The medical literature provides numerous cautionary tales of the dangers of an overreliance on single studies, no matter how comprehensive. One critique of meta-analysis (which often combines estimates from smaller and larger studies) is that the final conclusion (e.g., of whether a medical intervention is effective) sometimes differs from that reached in large-scale, controlled randomized trials ("megatrials"); the latter have historically been seen as the preferred "gold standard" (comparisons are summarized in Ioannidis et al. 1998, Lau et al. 1998). It has, however, been pointed out that the results of megatrials can differ from each other as much as they do from the pooled estimates derived from a meta-analysis (Furukawa et al. 2000; for case studies of this phenomenon in conservation biology see Chapter 26). Large-scale clinical trials draw on a homogenous pool of research subjects (a single species), use the same rigorous methodologies (double-blind trials), and have very large sample sizes (1000 to 10,000 subjects). It is therefore apparent that even when uniformity is maximized, there are still unknown sources of heterogeneity that affect the outcome of a treatment. This problem of study heterogeneity is likely to be far greater in ecology and evolutionary biology.

Given the obvious biological variation among ecosystems and species, and research budgets of ecologists and evolutionists that usually preclude sample sizes in the ten thousands, it is foolhardy to conclude too much from the outcome of any single study no matter how comprehensive it is. The long-term studies of the life histories and demography of red deer on the Isle of Rhum and Soay sheep in Scotland (Clutton-Brock and Coulson 2002), or the demographics of rainforest trees on the 50 ha plot in Panama (Condit et al. 1995) are model examples of the very best ecological data sets we have from single studies. Even these studies are, however, unlikely to estimate precisely the average strength of relationships if we want to build up a picture for a broader range of species or forest types. Worse still, due to temporal variation, even these studies have reported different effects depending on the time interval over which data were analyzed. For example, the effect of maternal body condition on offspring sex ratio was eventually shown to vary with population density in red deer (Kruuk et al. 1999), and demographic patterns vary among rainforest sites only short distances apart within Panama (Condit et al. 2005).

ARE WE REALLY SO DIFFERENT?

Preoccupation with exemplary studies can generate a mythical quest for the "ideal" study. Many researchers will flatly state that they conducted their study because previous tests of a hypothesis yielded contradictory outcomes. An unflattering but plausible interpretation of their statement is that they implicitly believe that the earlier studies were flawed, and a new study is required in which all confounding variables are controlled to obtain the true answer. If one is interested in making generalizations, this view is patently absurd. There is no one set of

conditions that create the perfect study, unless one wants to generate a hypothesis that is confined to one species, in one place, at a single point in time. This is a pointless task for ecologists and evolutionary biologists.

The myth of the "perfect study" creates a mindset where the main aim of a researcher might be to refute the findings of an earlier study that has gained prominence (i.e., it is thought to be "counter-intuitive"). Instead of trying to accumulate a body of evidence in the form of a population of effect sizes, researchers directly pit their results against those of other studies. This is most obvious whenever acrimonious disputes arise between research teams whose results lie on opposite sides of the $P = 0.05$ divide. Implicit in such a dispute is that one of the studies must be erroneous. These disputes are not confined to studies that have been closely replicated, and can arise even when studies are on different populations, species, or ecosystems (quasi replication sensu Palmer 2000).

Closer inspection of effect sizes often reveals that perceived conflicts among studies are illusionary. There is no reason for ecological studies with identical protocols to produce identical estimates of effect sizes (for a review, see Kelly 2006). First, average effect sizes and sample sizes in ecology and evolutionary biology are sufficiently low that sampling error can generate considerable variation in estimated effects (Møller and Jennions 2002). Summarizing using threshold *P*-values creates problems because studies that fail to refute the null hypothesis are treated as contradicting those that do (see Kalcounis-Rüppell et al. 2005). The absurdity of this is shown in Figure 23.2. Here the significant correlations in studies B and C are treated as being in agreement with each other, but in conflict with the findings from study A, even though studies A and B yielded very similar estimates of the effect size (Stoehr 1999). Low statistical power makes it likely that attempts to replicate a study that reported a significant effect will produce a nonsignificant result (Jennions and Møller 2003). Study D illustrates this argument graphically. If this study is replicated at the same scale, the confidence intervals will be similarly wide. Even if the estimated mean effect in Study D is the true mean, almost half the new estimates will be nonsignificant because they will fall to the left of study D with 95% confidence intervals that then overlap zero. It is instructive to view a graph plotting the function that relates the likelihood that a null hypothesis is rejected in an exactly replicated study, to the *P*-value obtained in an initial study; see Greenwald et al. (1996), or Figure 1 in Kelly (2006).

Before researchers embark on accounts of why their results differ from those of another study, they should first test whether sampling error alone provides a parsimonious explanation. In our experience, broad 95% confidence intervals in ecology and evolutionary biology mean that nonsignificant and significant results rarely differ more than expected by chance (i.e., effect size estimates overlap). But if two studies do differ, what should we conclude? If they are truly identical studies then this can *only* be due to chance. If they are not, then any of the innumerable

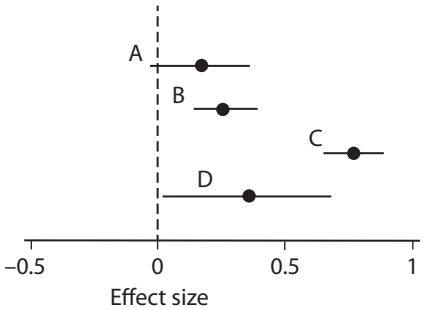

Figure 23.2. Effect sizes (means and 95% confidence intervals for Pearson's correlation coefficient) for four hypothetical studies. The null hypothesis (effect size = 0) is indicated by a vertical line.

factors that differ between them could account for the difference. We lack the data to decide which is the case because each study yields one effect size for a given hypothesis. This should be self-evident and seems trivial, but inspection of almost any recent journal will show that researchers still attempt to explain why two particular studies differ. Indeed, this is an almost compulsory component of the Discussion section of papers. Editors and reviewers will invariably ask authors to speculate as to why their findings differ from those of previous studies.

It is almost impossible to perfectly replicate biological studies, particularly those conducted in the field. This is because populations have different histories of selection and dispersal (Garant et al. 2005); genetic and environmental effects during development mean that individuals vary in their responses to identical stimuli (e.g., David et al. 2000); and responses often vary with time of year, physiological state, weather conditions, and so on (e.g., Qvarnström et al. 2000). Ecologists and evolutionists work with organisms, not atoms (clones or inbred lines in the laboratory are the closest analog we have). From this standpoint alone, researchers will achieve more for their field if they view their own work as contributing toward the ability to generalize, rather than seeing it as an attempt to validate or refute another result. Given modest statistical power, there is also limited ability even to reject a null hypothesis internal to their own study with high confidence, again suggesting that studies are best viewed as contributing to the wider picture.

HOW TO PRESENT RESULTS

Researchers who embrace meta-analysis can find it difficult to write Discussion sections of papers. Traditionally these are a forum where one must inflate the conclusions that can be drawn from a study in order to ensure its publication. When studies are viewed as contributing single data points to larger data sets, there is no incentive to overinterpret individual results. This can be daunting, but researchers should remind themselves of what they already know: their results are probabilistic. The outcomes of individual studies vary due to sampling error, and as a result of the genuine variation in the strength of causal factors under different conditions. No single study, no matter how large, is guaranteed to produce a universally correct answer (Ioannidis et al.1998). Pragmatically, we suggest that authors use the discussion section to highlight the accuracy of their estimates (i.e., sample sizes and the attention given to reducing measurement error); the quality of their experimental design (i.e., how well they have controlled for confounding variables); and the extent to which the tested relationships have been studied by others.

One suggestion is for authors to include small-scale meta-analyses that estimate the mean effect size for studies that have asked the same question(s) that is/are the focus of their own study; see Young and Horton (2005) for a formal requirement to do so in some journals. A comprehensive meta-analysis is a major task (Chapters 3 to 5). So, to make the task manageable authors can narrow their coverage to include only studies of the same species, taxonomic group, or ecotype. Even if the resultant meta-analysis is imperfect and does not conform to the strict protocols of a systematic review (Chapters 3 to 5; Roberts et al. 2006), it will still be of greater value than the current practice of citing a few papers that did or did not report a significant result in the same direction. Tables can be used, but if results are presented graphically (as in Fig. 23.2), the extent to which there are discrepancies between studies, and how power issues affect the likelihood of reporting a significant result (Colegrave and Ruxton 2003) is more easily grasped. In practice, there is precedence for this approach as one occasionally encounters primary research papers where the authors have tabulated "vote counts" of other studies that have asked the same question (e.g., Reynolds and Jones 1999). It is a small step to shift from reporting *P*-values to effect size estimates. Given the need for researchers to have incentives

for this extra effort, it is worth noting that publications that include this type of value-added quantitative information are more likely to be cited. They are a ready source of information for those compiling larger data sets to produce a more rigorous meta-analysis, and are usually formally acknowledged by then being cited.

META-ANALYSIS AND DECISIONS ABOUT STUDY REPLICATION

Once meta-analysis is embraced, "researchers see their piece of research as a modest contribution to the much larger picture in a research field" (Nakagawa and Cuthill 2007). From this perspective the goal is to ensure that multiple studies are available for subsequent analysis. This raises a question: How should studies be replicated? This answer is important because it influences the activities of individual researchers and of funding bodies. For example, is it better to fund a researcher who has reported an intriguing finding in an earlier pilot study and wishes to replicate the study with a larger sample size, or should funding be directed toward others who can ask the same question in other study systems (quasi replication)? One view is that some subdisciplines, such as behavioral ecology, have been damaged by a failure to precisely replicate studies that make exciting but controversial claims (for reviews, see Palmer 2000, Kelly 2006). If high-profile studies are false, they have undue influence in the long term (making them harder to dispel) but they can be dismissed quickly if they are rigorously scrutinized and swiftly replicated. Here we will make the counterargument that quasi replication is actually a more profitable approach for ecologists and evolutionists. There is, however, a caveat. This is only true *if* it is combined with a shift toward using the results of meta-analyses to guide the interpretation of published studies.

The appropriate form of study replication depends on how general we want to make our conclusions. We use a hypothetical case study to make our point. Consider the skepticism sometimes felt when a study produces an unexpected but impressive result. For example, it might be shown that feral cats with more symmetric whiskers sire more sons than daughters ($P = 0.02$, $n = 50$, $r = 0.33$; 95% CI: 0.04 to 0.62). This could lead to a plethora of "copy cat" studies (forgive the pun). The question of appropriate replication depends entirely on whether we are concerned that this specific study has miscalculated the biology of feral cats, or whether we want to test whether the finding is indicative of a wider phenomenon. We might be equally skeptical in both cases because a single study is the source of a whole new hypothesis (i.e., that paternal whisker symmetry predicts offspring sex ratios).

In disciplines like physics, researchers tend to accept that everyday studies produce "correct" results. The exceptions usually arise in areas at the forefront of theory where the requisite instrumentation or software is often at the limits of our technology. (For example, there is currently much debate about a recent result, using the CERN particle accelerator, indicating that muon neutrinos travel faster than the speed of light. Most physicists appear to assume that the speed was incorrectly measured because, if this is not the case, the finding would require major reorganization of established and well-corroborated theory). Greater confidence in the reliability of other researchers' findings is partly attributable to lower stochasticity and a better understanding of the effects of confounding variables (e.g., temperature, density) so that estimates of effect sizes tend to be more precise and replicable. In fields like physics greater attention can be given to testing whether a theory or phenomenon can be generalized (e.g., determining what classes of materials display superconductivity). Disinclination to repeat an original study might also be attributable to researchers having less affinity with knowing about a specific chemical compound rather than, say, learning about the general properties of a class of materials. In contrast, many biologists really want to know the truth about cats (or dogs or orchids). They

define themselves by the organisms they study, rather than the theories they are testing. In such cases, our hypothetical cat study will trouble these types of researchers if they believe that the pattern is spurious. They will remeasure the whisker symmetry-sex ratio relationship in other cat populations but, as we have already noted, it is impossible to replicate ecological studies with perfect precision. Unless researchers can demonstrate fraud, a flawed statistical analysis, or mismeasurement, it will not be possible to "disprove" the original study. There are too many uncontrolled factors. Eventually, however, cat researchers might accumulate sufficient studies to conduct a meta-analysis. If the weighted mean correlation across cat studies is close to zero we can, without making any direct judgments about the validity of the original study, conclude that it was unrepresentative.

We have discussed this example at length to make the following point: Does anyone really care that much about cats? Maybe not, at least when they are acting as scientists rather than pet owners. Close replication of a study is probably motivated more by the knowledge that, given current practices, it could become influential and be presented as "correct" simply because $P < 0.05$ without being independently verified. It might even become a textbook exemplar. One response is therefore to subject this single study to intense scrutiny. If it proves to have grossly miscalculated the average effect then there is scientific progress. It is rather limited progress though. In our case, we only end up knowing a lot more about cats. And what if the original cat study was a good estimate of the mean effect? A more economical approach is to test for a general rule. We can, for example, productively ask whether whisker (or other aspects of body) symmetry predicts sex ratios in felids by conducting studies on lions, cheetahs, pumas, and so on. If a meta-analysis indicates a weighted mean effect close to zero, we have learned that symmetry tends not to predict offspring sex ratios in felids. With hindsight, we can either infer that cats are unrepresentative of felids (which could then be tested) or that the original study reported a false positive. If there is a significant mean effect, however, our confidence that symmetry predicts sex ratios has been expanded to cover the average feline. Of course, in so doing we accept that there is less robust evidence available to confirm whether we have a good estimate of the predictive value of whisker symmetry in any given species. Generality trades off with specificity. The extent to which ecologists and evolutionary biologists can accept this trade-off is a major source of conflict between those who embrace and reject the use of meta-analysis.

The use of meta-analysis should ameliorate the view that quasi replication is uninformative about the validity of an earlier study. The only caveat is that we must ensure that quasi replication does not increase selective reporting compared to that for true replication (Palmer 2000, 470; e.g., truly replicated studies will be reported if they contradict the original study, while quasi-replicated studies that fail to detect an effect might go unpublished). There is currently a disinclination to reject the positive findings of an earlier study as inaccurate unless precise replication shows that they are anomalous (Kelly 2006). However, meta-analysis allows one to draw inferences about a specific study if one is prepared to generalize across disparate studies. For example, if a focal study reports a significant relationship, but meta-analysis of a wider range of studies shows that the mean effect is close to the null hypothesis, a meta-analyst would probably conclude (in the absence of additional information) that the results of the focal study were due to a type I error. In our case study, whisker symmetry in cats *might* have predictive powers that are not apparent in other felids, but if one takes a broader perspective the external evidence does not support this claim. Of course, if a specific study is an outlier with respect to the general distribution of effects, then it might be worthy of further investigation because it could have a genuine biological basis. Exactly the same population level approach can be taken to detect scientific misconduct if some authors consistently report larger effect sizes than

their coworkers, again with the caveat that some researchers might work on systems where measurement error is smaller, or tighter experimental designs are possible, or they might be more skilled at statistically controlling for confounding variables.

The points we have raised about study replication might seem trivial, but they are not. Embracing a meta-analytic perspective requires a profound cultural shift in how research findings are presented. If individual studies are given undue prominence then precise replication to challenge controversial findings will remain the favored response to controversial studies (Kelly 2006). Hopefully, however, greater use of meta-analysis will shift this mindset so that researchers are more interested in "the average study" rather than those at the extremes of the effect size distribution. Whether this will actually happen is still unclear. First, ecologists and evolutionary biologists are trained to observe and emphasize discontinuities in nature (e.g., to assign individuals to a species, or habitats to ecotypes). It requires a conceptual leap to switch to a worldview where, to pursue our case study one last time, one is comfortable speaking of symmetry predicting offspring sex ratio in felids, but can refrain from making a follow-up statement that this is true in feral cats but not in, say, lions if the single lion study had a value of $r = 0.10$ (95% CI: -0.19 to 0.39, $P = 0.50$, $n = 50$). The average felid is intangible, while lions and cats are real. Second, it would be amiss not to acknowledge that evolutionary biology (and perhaps ecology) is an unusual science because contingency and rare events matter. For example, although many patterns and rules have been detected that allow us to predict the direction of adaptive evolution (e.g., a shift in life history strategies toward earlier sexual maturation in response to predation; Roff 2002), sometimes a fascinating adaptation only evolves once. For example, one species of burrowing owl (*Athene cunicularia*) collects and places dung in front of its burrow (Levey et al. 2007). Experiments show that this behavior significantly increases the rate at which the owl's preferred prey of dung beetles are consumed. There is no way to generalize this result, because no other species use dung in this fashion, and the only replication possible is to validate the positive effect of dung placement on owl foraging success. In some respects, evolutionary biology is akin to economics where general laws can be formulated but rare events (which are common enough as a class) often lead to unique outcomes making predictions difficult (for a popular account of the "dismal sciences," see Taleb 2007).

THE ADVANTAGES OF "EFFECTIVE THINKING"

Meta-analysis and the use of effect sizes can improve ecological and evolutionary studies by allowing researchers to focus on questions that they could not previously answer either in practice or in principle. Nakagawa and Cuthill (2007) have coined the apt phrase "effective thinking" for the resultant mindset. Here are some advantages of "effective thinking":

(1) Power and wasteful explanations: It is a truism that sample sizes in ecology and evolutionary biology are small. For example, empirical studies that involve tracking the life histories of individuals to measure their reproductive success, growth rates, survival rates, or that attempt to estimate share of paternity using microsatellites, suffer severe logistic and funding constraints. Small sample sizes result in low statistical power and frequent failure to reject false null hypotheses. Although some ecological journals encourage the presentation of power analyses, this is still uncommon. Doing so voluntarily could punish researchers because, without a baseline reference for average power in a field (e.g., Jennions and Møller 2003), reviewers are more likely to reject papers when power appears low, say, $< 30\%$. Authors are therefore under pressure to discuss negative results as though they are conclusive. This is wasteful and generates spurious arguments. Presenting effect sizes and their confidence intervals (even though they convey

similar information) is a gentler way to remind readers about the extent to which they can draw inferences from specific tests (Colegrave and Ruxton 2003). In the long run, it also makes it easier to assess the repeatability of studies, by comparing the location and precision of effect size estimates (e.g., Fig. 23.2).

(2) Detecting trends: Given low statistical power, vote counting of significant studies is a very weak method to detect general biological trends. The use of effect sizes makes it far easier to detect patterns. In Figure 23.1, for example, if the criterion for significance was $P < 0.01$, only two of eight studies rejected the null hypothesis. Inspection of the graph shows, however, that all eight studies reported a positive relationship. This leads to a very different interpretation than that reached if one were to extract the eight *P*-values from the text of the paper (e.g., 0.001, ns, ns, ns, 0.001, ns, ns, ns). If the mean effect differs from the null value, deciding whether there is a causal relationship depends on the design of the original studies (see Chapter 24 for further discussion of how to interpret mean effect sizes).

(3) Future study design: Information about the average effect size can provide post hoc insight into why many published studies did not obtain significant results (e.g., due to low power to detect an average sized effect). It also ensures that future studies testing for the focal relationship in a specific context are designed with adequate statistical power. In addition, it creates the necessary benchmark against which comparisons can be made (see no. 5, below). However, there is a caveat; if there is a tendency for earlier studies to report inflated estimates of effect sizes (Chapter 15), then using these studies to design future work will lead to an overestimation of statistical power that, in turn, will increase the proportion of significant findings that are false positives (Ioannidis 2005c).

(4) Identifying sources of variation: Different studies, even those as precisely replicated as a biological system allows, rarely produce identical results. Compiling a data set of effect sizes allows us to ask why. We can first test whether the heterogeneity in effect size estimates is greater than expected by chance due to sampling error. If it is, then we have genuine conflict among studies and a new world to explore that was hidden when we only focused on significance testing in the original studies. The next step is to undertake exploratory studies to identify potential correlates of effect sizes. How these are interpreted depends on whether studies were randomly assigned with respect to the variables of interest. If they were (and this is often a judgment call), then we can tentatively posit a causal relationship between these factors and the relationship (effect size) under study. For example, if the correlation between body size and fecundity is stronger in deep water than shallow water marine species, we might causally attribute this to a depth effect. However, we cannot discount the possibility that the available species were drawn nonrandomly from the two habitats (e.g., it was harder to obtain data from larger bodied animals in deeper water). Also, because we have not experimentally manipulated depth, we cannot exclude the possibility that a correlate of depth (e.g., light levels or temperature) is responsible for the variation in effect sizes. Nonetheless, through judicial data exploration much progress can be made. For example, comparisons of effect sizes obtained from studies of colder and warmer waters at the same depth can corroborate or diminish an argument that temperature rather than depth affects the size-fecundity relationship. The use of such "natural experiments" is an unavoidable component of ecological and evolutionary research because some questions are simply not amenable to formal experimental manipulations.

The search for sources of variation in effect sizes is more likely to be important in ecology and evolution that in other areas of sciences. This is because there is greater variation in the range of study systems for which we want to draw general conclusions,

the methods used to collect data and test hypotheses are more variable, and our ability to control confounding variables (especially in field studies) is limited (Chapter 25). The use of meta-analysis models that include predictor factors or continuous covariates to explore variation in research findings is arguably one of the key advances that a shift to effect size thinking can deliver for ecology and evolutionary biology, and the synthesis of these two disciplines.

(5) Ranking the importance of factors: Within a single study researchers often test how well a range of factors (or experimental manipulations) predict changes in a response variable. If results are only reported in terms of *P*-values, it is not easy to rank their relative importance. In contrast, presenting effect sizes and measures of their variability offers a simple way for readers to compare the influence of different factors or treatments (see Fig. 24.9 in Chapter 24). The identical approach can be used to compile data from separate studies to identify which factors are strong or weak predictors, or to identify those factors where large confidence intervals for effect size estimates suggest that we need more data before we draw any conclusions. One could argue that when sample sizes are the same, *P*-values can be used to rank factors. This is true, but in ecology and evolutionary biology sample sizes are almost never identical, and may be consistently smaller for some variables because they are more costly or difficult to measure. For example, a sexual difference in body size is easier to measure than one in immune system effectiveness (one might also question whether it is reasonable to combine such different responses in a meta-analysis). A study of correlates of bib size in male sparrows by Nakagawa et al. (2007) is a nice case study illustrating how pooling effect sizes across studies can inform the direction of future research (see Fig. 24.8 in Chapter 24).

(6) Should I ask the same question? The use of cumulative meta-analysis allows us to test whether estimates of the mean effect size have stabilized (Chapter 15). If so, this implies that future studies of a similar nature are unlikely to meaningfully alter our conclusions. This encourages researchers to ask new questions, or to direct their attention to exploring finer-scale variation in the strength of an effect under different circumstances (see no. 4, above).

(7) Effect sizes as new variables: Effect sizes are themselves data points that can be used as either predictor or response variables in statistical analyses. We have already described their use as response variables whenever attempts are made to predict sources of heterogeneity in effect size estimates. The comparative method has been enormously effective in studies looking at the evolution of adaptive traits (Felsenstein 1985) and, to a lesser extent, in asking higher-level questions in ecology, such as those about community composition (e.g., Losos 1996, Cardillo et al. 2008). Comparative tests have led to major advances in our understanding of how traits coevolve and what drives the evolution of specific life histories and body shapes. Now that phylogenetic comparative analyses are becoming available for effect sizes (Chapter 17; Adams 2008, Felsenstein 2008, Lajeunesse 2009, Hadfield and Nakagawa 2010) we should see increased interest in studying patterns of coevolution between effect sizes and fixed traits or even between pairs of effect sizes. Researchers have asked why morphological traits like relative testes size are so much bigger for some species than others. (It is due to intense sperm competition in species where females mate with multiple males.) Equivalent questions can now be posed in the same way for more "dynamic" properties captured by effect sizes that once seemed less quantifiable, such as how boldness or shyness relate to fitness (Smith and Blumstein 2008), or the extent to which body size increases as temperature decreases (Adams and Church 2008). It is also worth remembering that, although still rarely done,

effect sizes can be used as predictor variables (Chapter 24). We can also ask questions about how effect sizes coevolve. For example, in species where mating with nonvirgin males has a more detrimental effect on female fecundity (Torres-Vila and Jennions 2005), are females more discriminating about mating with virgins? Effect sizes calculated using the proportion of females that choose virgins over nonvirgins in two-choice trials would allow this idea to be tested; that is, are the two effect sizes correlated?

(8) Improving meta-analyses: Reporting effect sizes in primary studies would greatly facilitate the extraction of effect sizes for meta-analyses. It would reduce the risk of transcription and calculation errors when compiling a data set for a meta-analysis, and would result in greater replicability for the meta-analyses based on these data.

(9) Management and Policy: In applied areas of ecology and evolutionary biology, those unfamiliar with the details of scientific methodology are often required to develop management strategies and formulate policies based on scientific findings. Reporting effect sizes is likely to reduce the likelihood that the potential magnitude of a given practice will be incorrectly estimated due to over-reliance on *P*-values.

CONCLUSIONS

In our view, publication practices in ecology and evolutionary biology overemphasize the value of individual studies. The resultant focus on *P*-values has led some researchers to believe their task is to confirm or refute isolated null hypotheses. However, on closer inspection this is almost never their real long-term goal. Even those who only seem interested in understanding their small corner of the natural world tend to have greater aspirations. No working biologist ever presents results in isolation. Invariably other studies, often on different species, taxa, or ecosystems, are cited. Why? Either there is an expectation that there is a general rule, so that studies detecting the same pattern or experiments identifying the same causal factor are cited; or the researcher thinks that his/her study differs from previous ones in a way that will influence causation, so that failure to obtain the same result is worth highlighting. Given this practice, even those ecologists and evolutionary biologists who are primarily interested in working out the details of their own study system should be happy to accept some responsibility for presenting data in a form that makes it easier to conduct meta-analyses.

We believe that the intellectual goal of most ecologists and evolutionary biologists is to uncover general rules in nature, and to identify the exceptions that push research in new directions. This goal is only achievable when we work on a scale that is larger than our own research projects. A science that seeks only to test an isolated hypothesis is merely a program to catalogue nature in a piecemeal fashion. Some empiricists have long accepted the reality that individual studies are small pieces of a big picture. In evolutionary biology, the advances in understanding that have come from the use of the phylogenetic comparative method perfectly illustrate this process. Biologists have learned to accept that grueling fieldwork is often boiled down to a single data point for a species in a phylogenetic regression. We should be equally comfortable with the fact that the real value of the statistical tests that we calculate is often not to confirm the occurrence of a phenomenon in our own study system (although this might be of great interest to ourselves and a few others), but rather to generate an effect size that can be pooled to explore trends at higher levels of analysis.

Finally, we should acknowledge that many of the points we have made in this chapter address issues that are beyond the immediate control of many meta-analysis practitioners. They lie in the domain of editors, funding agencies, and so on. Even so, today's young biologist

is tomorrow's chief editor or funding agency executive. This chapter is ultimately a work of advocacy that can hopefully be invoked by, for example, those querying editorial decisions or challenging the "conventional wisdom" of reviewers whose opinions are not always substantiated by valid quantitative analysis of the literature. It is worthwhile questioning current publication practices, because change does not occur without dissent and debate.